| Problem Chosen | 2025 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2513660 |

# Prediction Model of the Olympic Medal Tally


## Summary

The prediction of the Olympic medal table often attracts widespread attention. By establishing a multi-disciplinary combination prediction model based on multiple features, we achieve a relatively accurate prediction of the Olympic medal table.

For Task 1, we first build several **basic features** based on the disciplines. Using these features, we develop a **multi-disciplinary combination prediction model**, which successfully predicts the number of gold, silver, and bronze medals for each country at the 2028 Olympic Games. Compared to the 2024 Olympics, Slovakia, Fiji, and Turkmenistan are the countries most likely to see a significant rise in rankings, while the countries with the worst performance are Hong Kong, the Philippines, and Portugal.

Next, based on the prediction model, we identify four countries that are likely to win their first-ever Olympic medals at the next Olympic Games. By examining the features of countries that won their first medal in previous Olympics, we find that this is often attributed to participation experience. The four countries predicted in our model also follow this pattern, confirming the validity of the model.

Then, we constructed a **multiple linear regression model** to explain the impact of the number and type of Olympic events on the number of medals. At the same time, we developed a **simpler linear regression model** to obtain a more direct linear relationship.Furthermore, we use the major discipline advantage coefficient to measure the importance of each major discipline to a country. Additionally, in our study of the medal outcomes in sports events chosen by the host country, we were surprised to find a **radiating effect** from the host nation.

For Task 2, we visualize the "great coach" effect to verify its potential existence. Using both the **adjusted regression model** and the **SHAP-based XGBoost model**, we quantify the contribution rate of the "great coach" effect. The results show that it significantly influences only the gold and silver medals. Furthermore, we use the quantified "great coach" effect to determine the most suitable sports disciplines for investment in "great coaches" in countries like the United States, and estimate the impact of the "great coach" effect based on the increase in gold and silver medal scores.

For Task 3, we visualize **the most advantageous major disciplines** of each country on a map, discovering a consistency in North America’s advantageous disciplines. Additionally, using the K-means clustering algorithm, we classify the countries into three groups for visualization and find that the advantageous disciplines show a **"clustered" pattern**. Within each cluster, disciplines such as basketball and tennis exhibit a **strong correlation**.

Finally, we perform sensitivity analysis on the multi-disciplinary combination prediction model, assessing the fluctuations in predicted medal counts due to variations in the features.



**Keywords: Feature construction; Multi-disciplinary combination prediction model; SHAP; XGBoost model; Major discipline advantage coefficient**

# Contents

# 1 Introduction

## 1.1 Problem Background

The Olympic Games are the largest and most influential multi-sport event in the world, serving as an important venue for cultural exchange among nations. People from various countries often focus on individual events during the Olympics while closely watching the top-ranked countries on the final medal table. As a global celebration for all of humanity, the medal achievements of other countries also attract attention from all sectors of society.

**Figure 1: The Olympic Games**

Thus, predicting the final medal counts of each country at the Olympics has become a popular topic before every Olympic Games.

## 1.2 Restatement of the Problem

Based on the basic information and dataset provided in the problem statement, we need to address the following tasks:

- **Task 1**: Develop a model that can accurately predict the medal count for each country and evaluate the model's prediction accuracy. The model should be able to predict the medal table for the 2028 Los Angeles Olympics and identify the countries most likely to improve or decline in rankings. Additionally, the model should consider countries that have not yet won medals and be able to predict which countries will win medals for the first time in the next Olympics. Furthermore, the model should examine the relationship between Olympic events and the number of medals won by each country, identify the major disciplines most important to each country's medal count, and explore the impact of the host country's event choices on Olympic outcomes.
- **Task 2**: Provide evidence of the "great coach" effect on competition results and estimate the contribution of this effect to the final medal count. Additionally, select three countries and identify which Olympic disciplines they should consider investing in based on the "great coach" effect, and estimate the impact of this effect on the medal counts of these three countries.
- **Task 3**: Present other insights revealed by the model and provide the national Olympic committees with valuable information derived from these insights.

## 1.3 Our Work

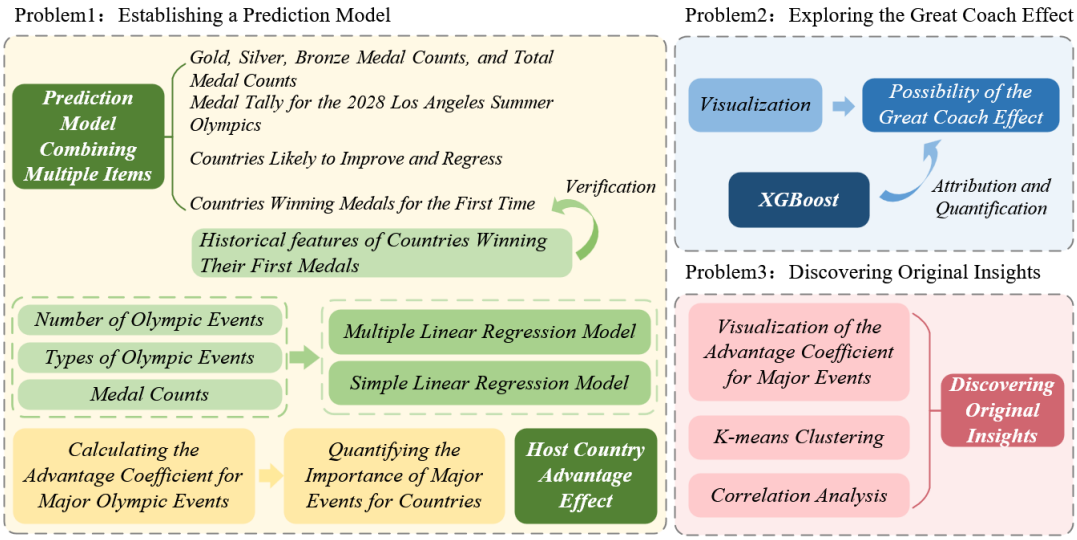


**Figure 2: Our Work**

# 2 Assumptions and Justifications

The prediction of Olympic medal counts is often influenced by many factors. Overly complex models not only require extensive computational resources but are also difficult to explain. Therefore, to simplify the modeling process, we make the following assumptions:

- **Assumption 1**: The medal counts from the past nine Olympic Games reflect the current athletic level of each country.
  **Justification**: Historical medal data from distant past Olympics may face issues such as country breakdowns and scheduling problems. Moreover, such data has limited practical significance for recent predictions.

- **Assumption 2**: The athlete data for the 2028 Olympics can be reasonably inferred from historical data.
  **Justification**: Since information about the athletes for the 2028 Olympics is unavailable, we must rely on historical data to make reasonable inferences, assuming that past data supports the validity of these predictions.

- **Assumption 3**: The coaching of a "great coach" always has a positive and beneficial effect.
  **Justification**: We assume that a "great coach" is an experienced coach who brings positive contributions to a team, without considering any potential negative impacts. Including negative effects would overly complicate the model.

- **Assumption 4**: A "great coach" specializes in a specific event or discipline, not a major or cross-disciplinary area.
  **Justification**: A "great coach" should be someone who excels in a specific subdiscipline, rather than someone who is broadly involved in multiple disciplines. This is more consistent with reality and would prevent unnecessary complexity in the model.

- **Assumption 5**: The contribution of a "great coach" to different events is treated equally.
  **Justification**: To avoid the complexity of modeling differences in influence,

such as the varying impact of a "great coach" on basketball versus volleyball, we assume a uniform contribution across events.

- **Assumption 6**: The differences in features due to the variations among events within a discipline are ignored, meaning a consistent feature construction approach is used to represent the medal count for each discipline.
  **Justification**: Avoiding excessive complexity due to minor distinctions between events within a discipline allows the model to remain manageable and focused on the key factors.

# 3 Notations

**Table 1: Notations used in this paper**

| Symbol | Description |
|---|---|
| $x_1$ | Traditional Advantage in a Discipline |
| $x_2$ | Potential Advantage in a Discipline |
| $x_3$ | Discipline Participation Rate |
| $x_4$ | Discipline Fit |
| $x_5$ | Athlete Gender Ratio |
| $x_6$ | Star Athlete Effect |
| $\lambda$ | Host Country Effect |

# 4 Data preprocessing

The data records provided by the International Olympic Committee (IOC) span several years, resulting in a significant amount of data that requires cleaning. This includes, but is not limited to, issues such as country dissolution, abnormal event names, and duplicate data. Therefore, we begin the data cleaning process with the following steps:

- **Program Table**:
  - Delete data before 1992 (excluding 1992).
  - Remove events that were not held in the past nine Olympic Games (1992–2024).
  - Replace missing secondary event names with the names of primary events.
  - Replace missing secondary event abbreviations with the corresponding secondary event abbreviations.
- **Athletes Table**:
  - Delete data before 1992 (excluding 1992).
  - Change full names of secondary events to their corresponding abbreviations.
  - Standardize the names of tertiary events across different years.
  - Remove special athlete data (e.g., athletes who do not represent a country, refugees competing under the Olympic flag, athletes from countries that have since dissolved, or countries that have never competed in future Olympics).
  - Remove data for countries that no longer exist or have dissolved.
  - Standardize country names across different tables.

# 5 Feature-Based Multi-Disciplinary Combination Prediction Model

For Task 1, if we were to build a separate prediction model for each country, this would involve a massive workload. Moreover, for every future prediction, all the country's models would need to be adjusted again, which is clearly unfavorable for prediction tasks. Therefore, we aim to develop a universally applicable model for all countries. This model should not only effectively predict medal counts but also offer excellent interpretability.

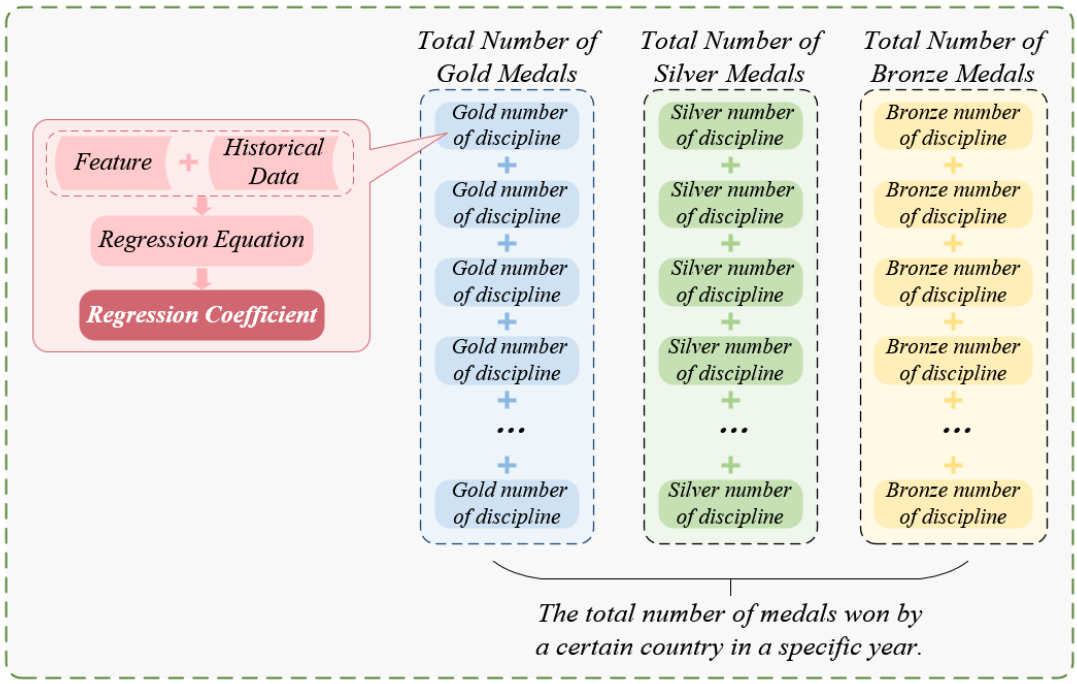


**Figure 3: Feature-Based Multi-Disciplinary Combination Prediction Model**

To address this, we established a hierarchical feature-based multi-disciplinary prediction model. This model consists of multiple regression fitting equations for various disciplines, taking into account factors such as the characteristics of athletes participating in the predicted Olympics and quantifying the potential impact of historical data on the predicted medal counts. The following is an example of feature construction for gold medal prediction in one specific discipline.

## 5.1 Feature Construction

Feature construction is based on the various sub-disciplines within the Olympic major disciplines. The first consideration should be the characteristics of the sub-disciplines themselves. For instance, whether a country has a traditional advantage in a particular sub-discipline can be considered as a general influence. At the same time, special attention should be given to the athlete effect, which refers to the impact of the characteristics of athletes in different Olympic editions on medal counts, representing an individual influence. Furthermore, the host country effect should also be taken into account, as it reflects an external influence brought by the Olympic host nation.

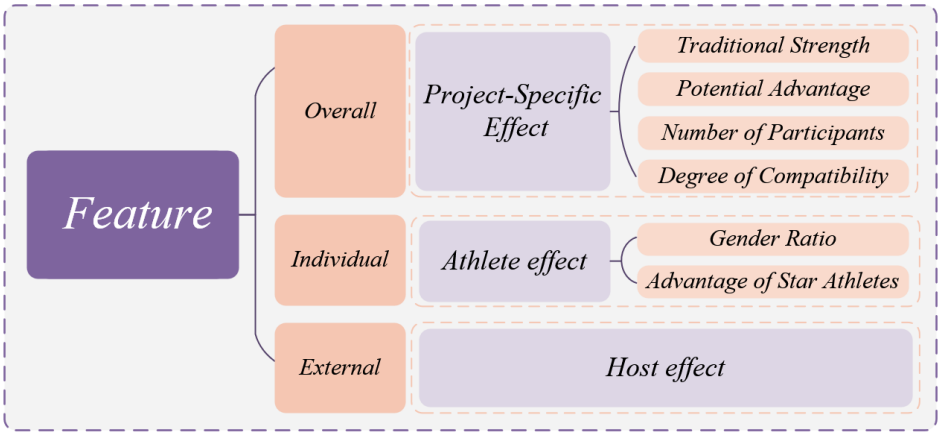


**Figure 4: Feature Construction Diagram**

### 5.1.1 Discipline Effect

- **Traditional Advantage in a Discipline** $x_1$

If a country has consistently performed well in a discipline over previous Olympic Games, it can be considered to have a traditional advantage in that discipline. This advantage is defined as the total score of gold, silver, and bronze medals earned in that discipline over the past 9 years, with 3 points assigned to a gold medal, 2 points to a silver medal, and 1 point to a bronze medal.

$$x_1 = \sum_{i=1}^{10} \sum_{j=1}^{n} \sum_{m=1}^{3} k c_m \tag{1}$$

In Equation 1, $n$ represents the number of events in the discipline, $k$ is the scoring coefficient, and $c_m$ denotes the number of medals.

- **Potential Advantage in a Discipline** $x_2$

Potential Advantage in a Discipline a country may not have a strong traditional strength in a discipline, but it may possess potential for development. This characteristic can be simply defined as the average growth in the total score of medals in the discipline (reflecting progress potential), combined linearly with the total number of appearances in the events within the discipline (reflecting competition experience).

$$x_2 = \overline{\Delta s} + \sum_{i=1}^{n} t_n \tag{2}$$

In Equation 2, $\overline{\Delta s}$ represents the average growth value, and $t_n$ denotes the total number of appearances by a country in an event within a discipline

- **Discipline Participation Rate** $x_3$

The total number of participants from a country in a discipline.

- **Discipline Fit** $x_4$

When the number of events in the Olympic Games is more concentrated in the disciplines that a country excels in, the country is relatively in an advantageous position.

The discipline fit is defined as

$$x_4 = \frac{s_i}{s_t} u \tag{3}$$

In Equation 3, $s_i$ represents the medal score of the discipline, $s_t$ is the total medal score of the country, and $u$ denotes the total number of events in the discipline.

### 5.1.2 Athlete Effect

- **Athlete Gender Ratio** $x_5$

The male-to-female ratio of athletes in a discipline may impact the number of medals, so the proportion of male athletes in the total number of athletes is calculated.

- **Star Athlete Effect** $x_6$

Star athletes with previous medal experience are often expected to win more medals. Athletes who have won gold, silver, and bronze medals in the Olympics are assigned 3, 2, and 1 points, respectively, while those who have not won a medal are assigned 0 points.

$$x_6 = \sum_{i=1}^{n} k a_g \tag{4}$$

In Equation 4, $k$ represents the scoring coefficient, and $a_g$ denotes the number of athletes who have won a gold medal.

### 5.1.3 Host Country Effect $\lambda$

As the host country of the Olympic Games, a nation has a home-field advantage. The countries that have hosted the Olympics in the past nine years (United States, Australia, Greece, China, United Kingdom, Brazil, Japan, France, United States) are analyzed. The medal performance of these countries is compared when they were the host country versus when they were not.

When a country serves as the host country

$$\gamma_1 = \frac{The\ total\ number\ of\ gold\ medals\ in\ the\ discipline}{The\ total\ number\ of\ gold\ medals\ across\ the\ past\ 9\ OlympicGames} \tag{5}$$

And when it is not the host country, the average medal performance for each country is calculated

$$\gamma_2 = \frac{The\ total\ number\ of\ gold\ medals\ in\ the\ discipline}{The\ total\ number\ of\ gold\ medals\ across\ the\ past\ 9\ OlympicGames} \times \frac{1}{9} \tag{6}$$

The difference between these two values is denoted as

$$\lambda = \overline{\gamma_1} - \overline{\gamma_2} \tag{7}$$

The difference $\lambda$ can be considered as the host country effect. This effect applies only to the host country. For the 2028 Olympic Games, it is sufficient to add this effect to the final medal predictions for the United States.

### 5.1.4 Normative Normalization

Due to the varying scales of the features, directly using them for regression fitting

makes it difficult to assess the importance of each feature based on the coefficients. Therefore, all features are normalized using the norm.

$$x_i' = \frac{x_i}{\sqrt{\sum_{i=1}^{6} x_i^2}} \tag{8}$$

$x_i$ represents the constructed feature value, and $x_i'$ denotes the normalized feature value.

## 5.2 Multi-Discipline Combined Prediction Model

### 5.2.1 Establishment of the Multi-Discipline Regression Combination Model

Based on the basic features of the disciplines mentioned above (excluding the host country effect), a multiple linear regression model is established for predicting the number of gold medals in each discipline, as follows:

$$y_i(t) = \sum_{i=1}^{6} a_i x_{\mathrm{i}} + \beta_0 \tag{9}$$

$y_i(t)$ represents the predicted number of gold medals for the target year, and $\beta_0$ is the constant term. Considering the impact of historical data on the number of gold medals won, a lag term is included, as follows:

$$y_i(t) = \sum_{i=1}^{6} a_i x_{\mathrm{i}} + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \tag{10}$$

The sum of the three-year lag terms in Equation 10 can be considered a comprehensive reflection of a country's past sports performance. It is the final result influenced by multiple factors.

Furthermore, by summing the results of each discipline, we obtain the total number of gold medals for a country.

$$y_g(t) = \sum_{i=1}^{n} \left( \sum_{i=1}^{6} a_i x_{\mathrm{i}} + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \right) \tag{11}$$

In the equation 11, $n$ represents the total number of disciplines in the prediction year. Similarly, we can construct the predicted number of silver medals $y_s(t)$ and bronze medals $y_b(t)$. The total number of medals is

$$y_{total} = y_g(t) + y_s(t) + y_b(t) \tag{12}$$

The regression coefficients for discipline $y_i(t)$ are fitted using data from that specific discipline for each country. This is because we are more interested in the impact of features on the same discipline. For example, we use the gold medal data for the Athletics discipline in Table 2.

**Table 2: Fitting Results for the Athletics Discipline**

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $\beta_0$ | $y(t-1)$ | $y(t-2)$ | $y(t-3)$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 16.73 | 4.34 | -1.17 | 0.07 | 0.13 | 8.68 | 0 | 3.15 | 1.36 | 8.32 | 0.92 |

The regression fit for $R^2$ is 0.92, indicating an excellent fit. The fitting coefficients for other disciplines are also above 0.85. For the Athletics discipline, the influence of traditional strengths is significant. This suggests that the teams that typically compete for gold medals in athletics are usually the traditional powerhouses.

### 5.2.2 Prediction of Regression Coefficients and Handling of Missing Data for the Prediction Year

Using the discipline regression combination model from the previous section presents two major challenges. First, when the number of medals for the prediction year is unknown, the regression coefficients are also unknown. Second, the lack of athlete records for the 2028 prediction year results in missing feature values.

When predicting medal counts directly, the complexity of factors influencing medal counts, such as personnel changes, makes it difficult to accurately predict future outcomes using historical data. Therefore, time series prediction has its limitations. However, for the regression coefficients of features, which are single-factor variables, time series prediction can effectively capture their temporal characteristics.

We use the ARIMA model to predict the feature coefficients for the past nine years. For example, we can use the Swimming discipline.

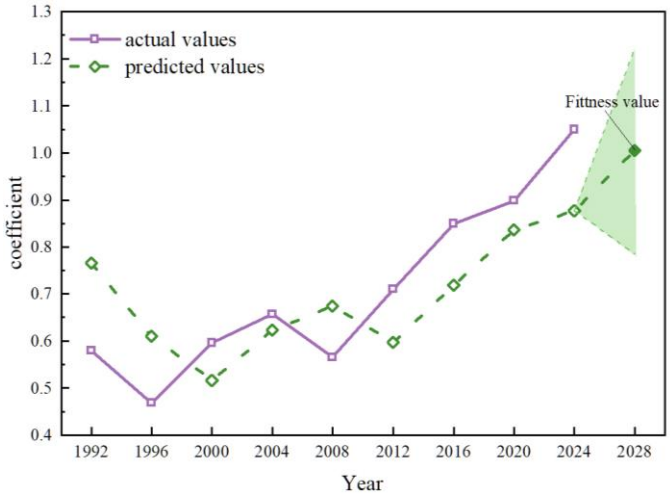


**Figure 5: ARIMA Regression Coefficient Prediction**

Additionally, the handling of missing feature data for the 2028 prediction is shown in Table 3.

**Table 3: Handling of Missing Feature Data**

| Discipline Traditional Strength | Discipline Potential Strength | Number of Participants in Discipline | Fit Degree of Discipline | Gender Ratio in Discipline | Star Athlete Advantage |
|---|---|---|---|---|---|
| Known fixed value | Known fixed value | Weighted average of the last three years | Calculated from 2028 event data | Weighted average of the last three years | Calculated from 2028 event data |

For the number of participants and gender ratio in the discipline, the weights are assumed to follow an exponential decay, with values of 0.6, 0.25, and 0.15 for the most recent, second most recent, and third most recent years, respectively.

### 5.2.3 Model Limitations and Improvements

The multi-discipline combination prediction model established above cannot predict medals for entirely new disciplines that have never appeared in the past nine Olympics. Therefore, we need to consider the medal allocation for these completely new disciplines.

Since the new disciplines are proposed by the host country, they may tend to favor the host country. To address this, we first calculate the percentage of gold medals won by the host country in the new disciplines, denoted as $\alpha$, which reflects the average performance of the host country in these new disciplines. Since the 2028 host country is the United States, we separately calculate the average performance of the United States as the host country in previous years (the U.S. has hosted the Olympics four times), denoted as $\beta$.

Both $\alpha$ and $\beta$ influence the final results for the completely new disciplines, but their impact levels are different.

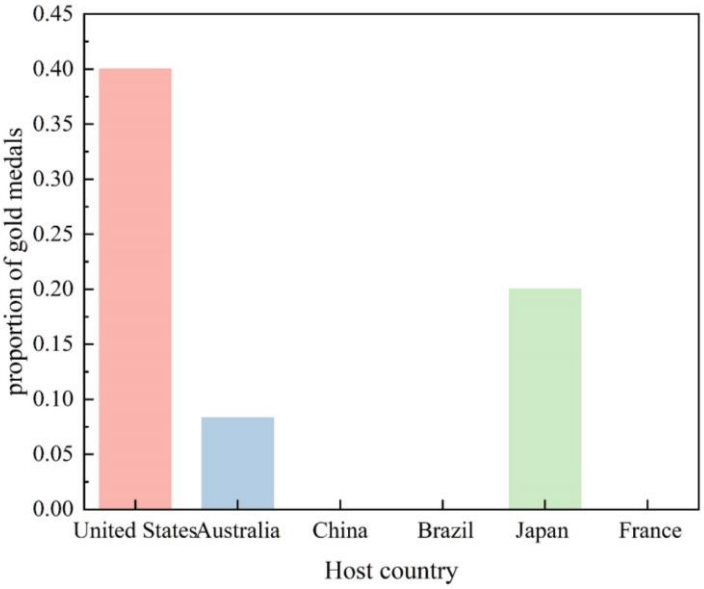


**Figure 6: Comparison of Performance in New Disciplines Added During the Last Nine Olympics Hosted by the US and Other Countries**

When the United States is the host country, its gold medal win rate is significantly higher than that of other host countries.

Therefore, we can estimate the number of gold medals the United States will win in the completely new disciplines as follows:

$$N = m(0.3\alpha + 0.7\beta) \tag{13}$$

In Equation 13, $N$ represents the estimated number of gold medals the United States will win in the completely new disciplines. $m$ is the total number of gold medals available in these completely new disciplines, which is 8 (Lacrosse, Cricket, Squash, and Flag Football each have 2 events, while Baseball/Softball is not considered a completely new discipline because it has appeared multiple times in the past decade and can be handled by the model). Calculating and rounding, we get $N = 3$.

## 5.3 Medal Predictions for the Los Angeles Summer Olympics

We use our multi-discipline combination prediction model to forecast the medal counts for the 2028 Los Angeles Summer Olympics. Below are the predicted gold, silver, and bronze medal counts, as well as the total medal counts, for the top seven countries in the medal table.

**Table 4: 2028 Olympic Medal Predictions**

| NOC | Gold | Silver | Bronze | Total | 2028 Rank | 2024 Rank | Change |
|---|---|---|---|---|---|---|---|
| USA | 50 | 56 | 32 | 138 | 1 | 1 | 0 |
| China | 34 | 40 | 21 | 95 | 2 | 2 | 0 |
| Japan | 26 | 13 | 8 | 47 | 3 | 3 | 0 |
| Great Britain | 23 | 22 | 9 | 54 | 4 | 7 | 3 |
| Australia | 17 | 15 | 9 | 41 | 5 | 4 | -1 |
| France | 16 | 20 | 14 | 50 | 6 | 5 | -1 |
| South Korea | 13 | 9 | 6 | 28 | 7 | 8 | 1 |

Looking at the top-ranked countries, the top three in 2028 remain the same as in 2024: the USA, China, and Japan. However, Japan shows a trend of closing the gap with China. Great Britain, Australia, and France still rank highly, but their positions have changed significantly, with Great Britain showing the most improvement.

Notably, the Netherlands, which ranked sixth in 2024, is predicted to fall out of the top seven in 2028, dropping to 12th place.

**Table 5: 2028 Olympic Medal Prediction Intervals**

| NOC | Gold | Range | Silver | Range | Bronze | Range | Total | Range | 2028 Rank |
|---|---|---|---|---|---|---|---|---|---|
| USA | 50 | [47,53] | 56 | [51,61] | 32 | [29,36] | 138 | [127,150] | 1 |
| China | 34 | [32,36] | 40 | [36,44] | 21 | [19,23] | 95 | [87,103] | 2 |
| Japan | 26 | [24,27] | 13 | [12,14] | 8 | [7,9] | 47 | [43,50] | 3 |
| Britain | 23 | [22,25] | 22 | [20,24] | 9 | [8,10] | 54 | [50,59] | 4 |
| Australia | 17 | [16,18] | 15 | [14,16] | 9 | [8,10] | 41 | [38,44] | 5 |
| France | 16 | [15,17] | 20 | [18,21] | 14 | [12,15] | 50 | [45,53] | 6 |
| South Korea | 13 | [13,14] | 9 | [8,10] | 6 | [5,7] | 28 | [25,31] | 7 |

The main source of prediction error in the model comes from the estimation of regression coefficients. To estimate the maximum fluctuation range of the medal predictions, we calculate the minimum and maximum values of the confidence intervals for the feature coefficients of each discipline. These values are used as the error range for the medal predictions for each discipline. The results of these interval calculations are shown in Table 5.

We observe that the gold medal prediction intervals have the smallest fluctuation range, while the silver and bronze medal intervals show larger variations. However, these variations do not affect the final rankings of the countries. Therefore, our prediction model is highly reliable.

From Table 6, we find that Slovakia, Fiji, and Turkmenistan are the three countries with the most improvement.

**Table 6: Top Three Countries with the Most Improvement**

| NOC | Gold | Silver | Bronze | 2028 Rank | 2024 Rank | Change |
|---|---|---|---|---|---|---|
| Slovakia | 1 | 1 | 0 | 42 | 83 | 41 |
| Fiji | 2 | 0 | 0 | 34 | 73 | 39 |
| Turkmenistan | 2 | 2 | 1 | 28 | 63 | 35 |

**Table 7: Top Three Countries with the Greatest Decline**

| NOC | Gold | Silver | Bronze | 2028 Rank | 2024 Rank | Change |
|---|---|---|---|---|---|---|
| Hong Kong | 0 | 0 | 4 | 76 | 36 | -40 |
| Philippines | 0 | 1 | 3 | 67 | 36 | -31 |
| Portugal | 0 | 0 | 3 | 79 | 49 | -30 |

And Hong Kong, the Philippines, and Portugal are the three countries with the greatest decline, as shown in Table 7.

## 5.4 Prediction of Countries Winning Their First Medals

Based on the medal table data predicted by the model for 2028, there are a total of four countries that are expected to win their first medals in 2028.

**Table 8: First-Time Medal-Winning Countries**

| Country | Discipline | Medal Type |
|---|---|---|
| Papua New Guinea | Weightlifting | Silver |
| Haiti | Athletics | Bronze |
| Seychelles | Swimming | Bronze |
| Nepal | Shooting | Bronze |

Notably, Papua New Guinea is expected to win its first medal, a silver, in the Weightlifting discipline at the 2028 Olympics. The other three countries are also predicted to win their first bronze medals.

Since the model predicts the number of medals, the error intervals for each discipline do not significantly affect the medal type (because the results are rounded). Therefore, the model's predictions are accurate.

In addition to using the model to predict medal counts, we should also pay special attention to the characteristics of countries that won their first medals in previous Olympics.

We analyzed the disciplines in which countries won their first medals in each Olympic Games. Based on the features constructed earlier, we calculated six feature values for each discipline and took the average of these features. The goal is to identify common characteristics among the medal-winning disciplines.

Additionally, we calculated the average feature values for the corresponding disciplines of countries that did not win medals. This helps us identify common characteristics among the non-medal-winning disciplines. Over the past nine Olympics, we found nine sets of corresponding feature comparisons. We used average values to reduce errors.

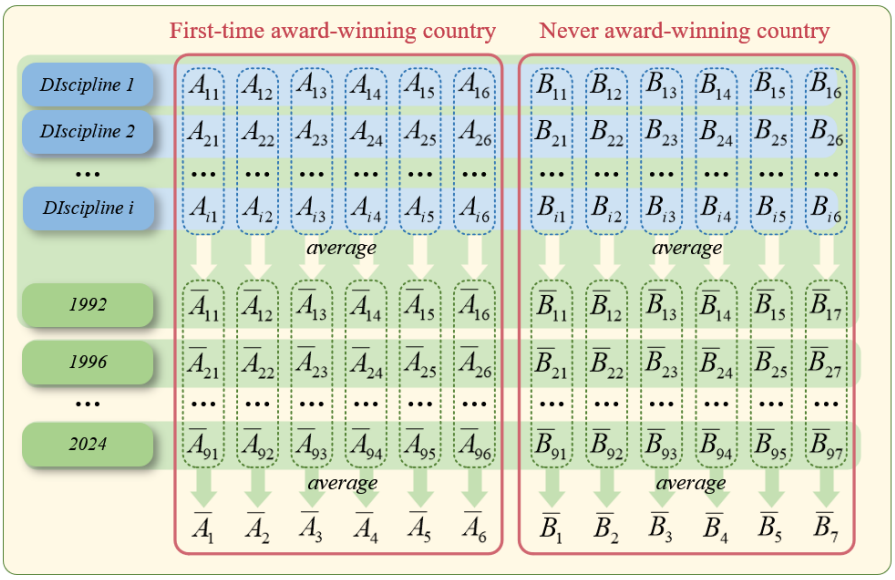


**Figure 7: Feature Analysis of Medal-Winning and Non-Medal-Winning Disciplines**

In Figure 8, the medal-winning disciplines, compared to the non-medal-winning disciplines, do not have traditional discipline advantages, high fit degrees, or star athlete effects.

The key distinguishing features between medal-winning and non-medal-winning disciplines are a country's potential advantage in the discipline and the number of participants. These factors are mainly attributed to the experience gained from multiple participations. The gender ratio in the discipline has some influence but is not a major factor.

The average values for the disciplines of the countries predicted to win their first medals in 2028 also follow this pattern. This confirms the validity of our predictions.

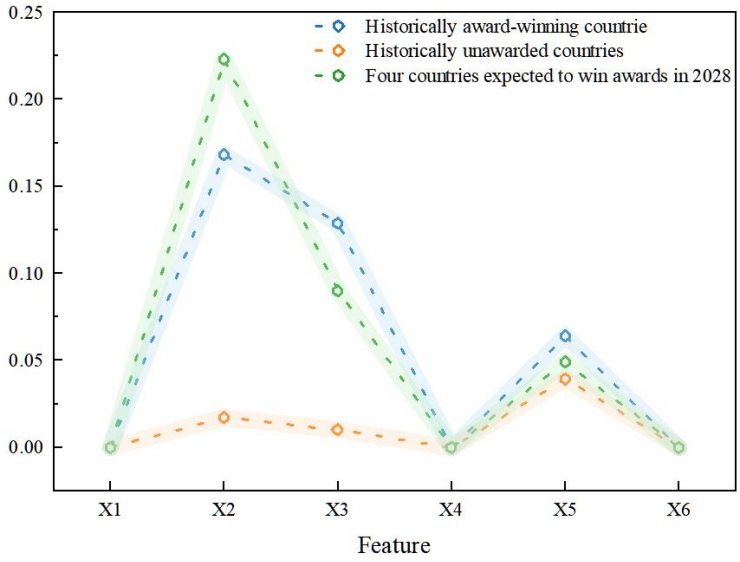


**Figure 8: Comparison of Characteristics Between Medal-Winning Disciplines and Non-Medal-Winning Disciplines**

## 5.1 The Selection of Events and Its Association with Medal Counts

### 5.1.1 relationship between the events and medals

In the previous features, we constructed a feature to measure the alignment of disciplines. This feature assesses the relationship between the type and number of events and the number of medals. For example, it can be defined in terms of gold medals as follows:

$$y_g(t) = \sum_{i=1}^{n}\left(\sum_{i=1}^{3} a_i x_i + a_4 \frac{s_i}{s_t} u + \sum_{i=5}^{6} a_i x_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0\right) \tag{14}$$

In Equation 14, $y_g(t)$ represents the total number of gold medals for a country. $s_i$ is the alignment score for the disciplines. $s_t$ is the total score for the country. $dd$ is the total number of events within the disciplines.

Although this equation provides a quantitative relationship, it combines too many other features. To more clearly show the relationship between the number and type of events and the number of medals, we define

$$\theta = \sum_{i=1}^{m} \frac{s_i}{s_t} u \tag{15}$$

We define this as the alignment of a country with the Olympic events for that year, where $m$ is the total number of disciplines.

We then take the average alignment over the past nine Olympics and regress it against the number of gold medals, resulting in Figure 9.

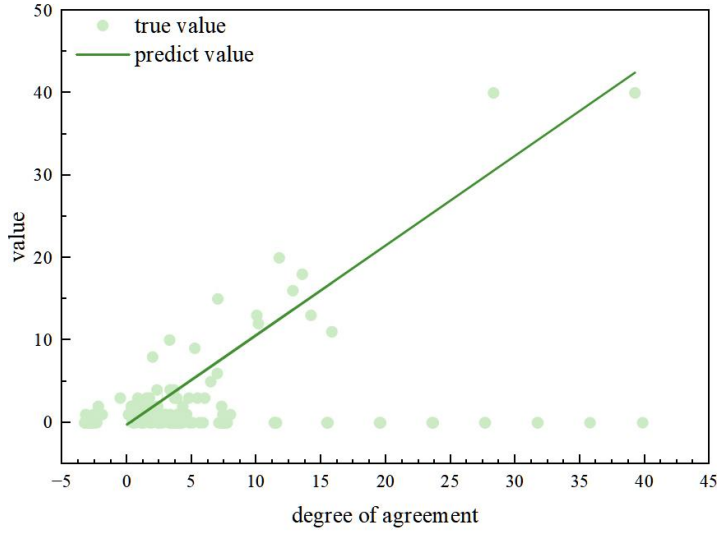


**Figure 9: Regression of Alignment with Gold Medal Counts**

This results in a simpler linear relationship between the two

$$y_g(t) = 1.09 \sum_{i=1}^{m} \frac{s_i}{s_t} u - 0.239 \tag{16}$$

The regression coefficient $aa$ is 0.85, indicating a good fit.

### 5.1.2 The Most Important Sport for a Country

The most important sport for a country in the Olympics should be the one in which the country wins the most medals. It is essential to consider not only gold medals but also silver and bronze medals. We assign 3 points for a gold medal, 2 points for a silver medal, and 1 point for a bronze medal. This leads to the sport advantage coefficient:

$$\eta = \frac{M_i}{M_t} \tag{17}$$

In Equation 17, $M_i$ is the medal score for the sport, and $M_t$ is the total medal score for the country.

The higher the advantage coefficient, the more important the sport is for the country. We will use the United States and China as examples.

**Table 9: The Most Important Sports for the US and China**

| US Advantage Sport | Advantage Coefficient | China Advantage Sport | Advantage Coefficient |
|---|---|---|---|
| Aquatics | 0.36 | Aquatics | 0.19 |
| Athletics | 0.20 | Badminton | 0.18 |
| Skateboarding | 0.07 | Table Tennis | 0.14 |

We can see that the most important sport for both the US and China is Aquatics, but the importance levels differ. Aquatics is much more significant for the US than for China. This also reflects that China's advantage is less pronounced, with a more balanced performance across different sports.

### 5.1.3 The Impact of Host Country's Selected Events on Medal Outcomes

We have already provided some explanation of the host country's event selection impact in Figure 6.

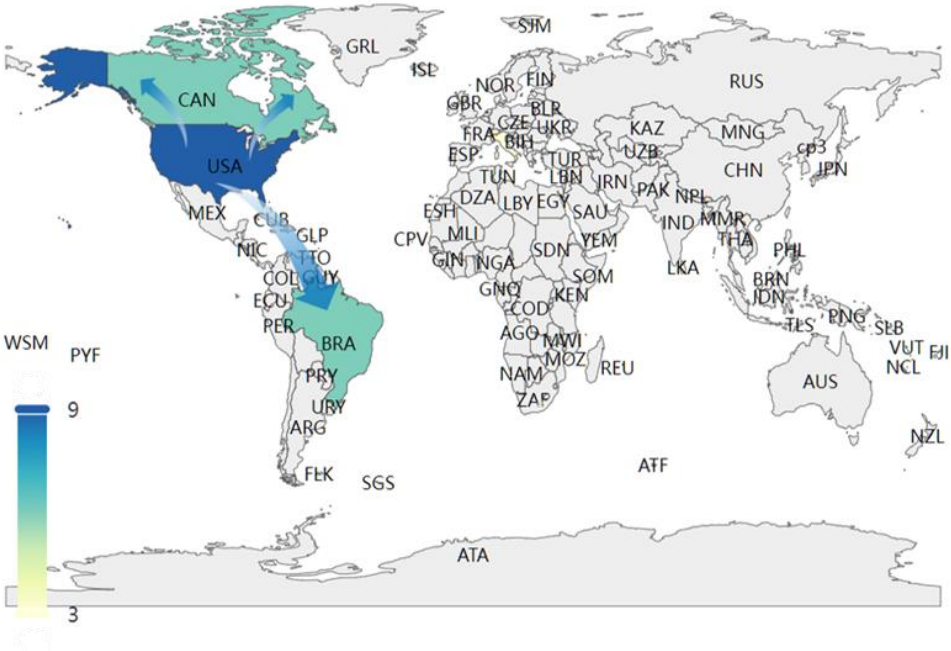


**Figure 10: World Map Distribution of Scores for New Disciplines**

However, we discovered another interesting phenomenon: the host country's spillover effect.

For the new events chosen by the host country, neighboring countries often achieve success in those events as well. This suggests that cultural influence within the same region may play a role.

For example, in the 1996 Olympics hosted by the United States, three new disciplines were added: Mountain Bike, Softball, and Beach Volleyball. We recorded the total medal scores for these three disciplines for each country and plotted them on a world map.

The results show that the events chosen by the host country have a certain "spill-over" effect. The countries that win medals in these events are not too far from the host country.

# 6 The "Great Coach" Effect

## 6.1 Evidence of the "Great Coach" Effect

We will use the coaching careers of Lang Ping and Béla Károlyi as examples to explore the evidence of the "Great Coach" effect.

Since "great coaches" often specialize in a specific discipline or event, using only the number of medals for analysis is too broad. Therefore, this section uses detailed scores for visualization.

By comparing the total scores before and after Lang Ping's coaching, it is clear that the total medal scores generally increased when she was coaching. The only exception was in 2020, where there was a decline.

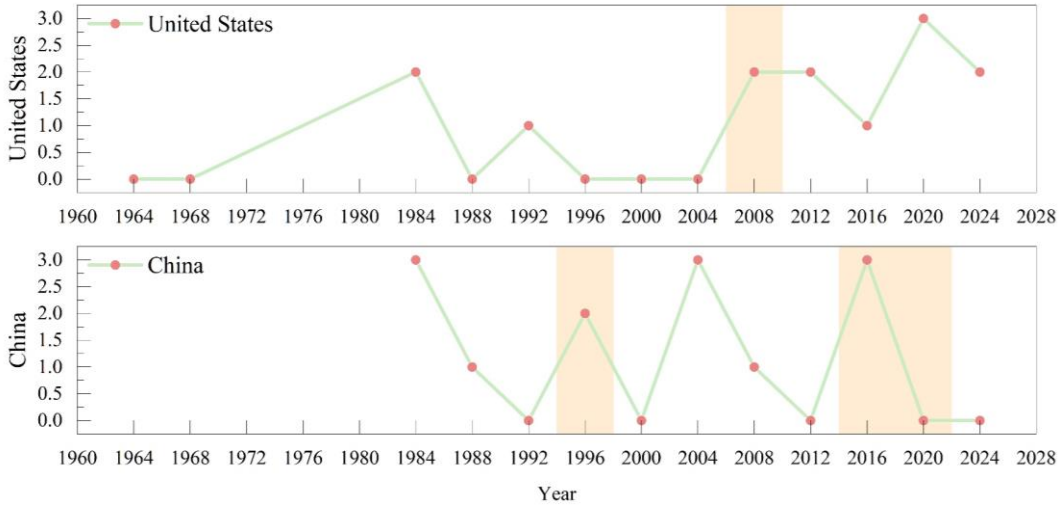


**Figure 11: Medal Score Changes During Lang Ping's Coaching Years**

When Béla Károlyi first coached in Romania and the United States, there was a significant increase in total scores. However, subsequent changes did not consistently show the "Great Coach" effect. This is because Béla Károlyi's career was more complex; he was not always the head coach of a national team, making the "Great Coach" effect less pronounced. In contrast, Lang Ping clearly demonstrated the "Great Coach" effect.

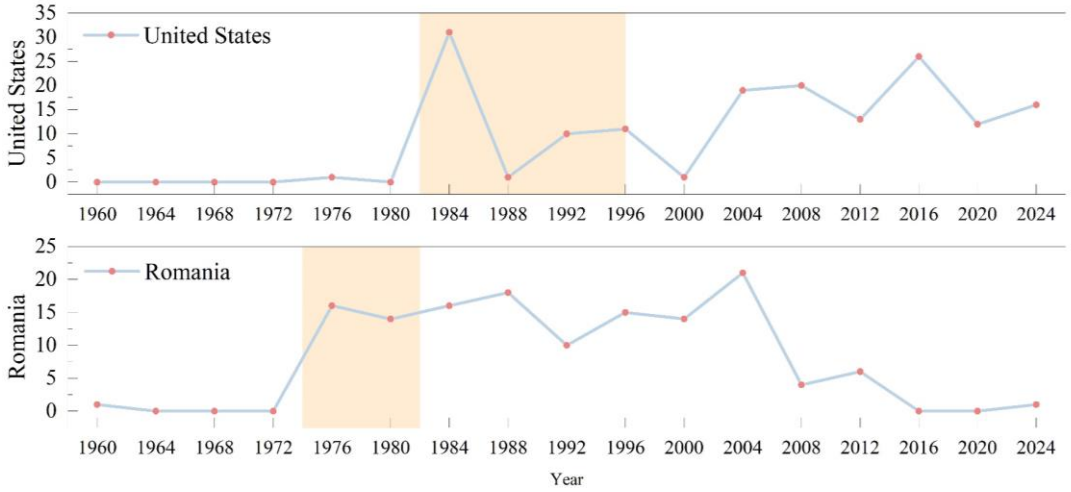


**Figure 12: Medal Score Changes During Béla Károlyi's Coaching Years**

## 6.2 Quantifying the Contribution of the "Great Coach" Effect

### 6.2.1 Quantification Using an Adjusted Regression Model

We make similar modifications to the previous features to make them applicable to events (since the alignment of disciplines is not suitable for events, we consider only the other five features). Next, we establish a similar prediction model based on events and use it to predict the years in which each "great coach" was involved in coaching.

$$\begin{cases} y_a(t) = \sum_{i=1}^{5} a_i r_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \\ y_b(t) = \sum_{i=1}^{5} a_i r_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \\ y_c(t) = \sum_{i=1}^{5} a_i r_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \\ y(t) = k_1 y_a(t) + k_2 y_b(t) + k_3 y_c(t) \end{cases} \tag{18}$$

$y_a(t)$、$y_b(t)$、$y_c(t)$ represent the number of gold, silver, and bronze medals, respectively. Let $r_i$ denote the event feature, $k_1$、$k_2$、$k_3$ the scoring coefficient, and $y(t)$ the total score.

The difference between the predicted values and the actual values can be attributed primarily to the "Great Coach" effect. We have:

$$r_6 = y(t) - \hat{y}(t) \tag{19}$$

Therefore, for each year in which a "great coach" was involved in coaching, we can obtain the event feature and the deviation caused by the "Great Coach" effect.

We then use these features and the deviation values to construct a new regression model.

$$\begin{cases} y_\alpha(t) = \sum_{i=1}^{6} b_i r_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \\ y_\beta(t) = \sum_{i=1}^{6} b_i r_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \\ y_\gamma(t) = \sum_{i=1}^{6} b_i r_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \\ y(t) = k_1 y_\alpha(t) + k_2 y_\beta(t) + k_3 y_\gamma(t) \end{cases} \tag{20}$$

Using the data from each coaching year of Lang Ping and Béla Károlyi, we fit a regression model. The regression coefficients for the event features and the "Great Coach" effect are shown in Table 10.

**Table 10: Regression Coefficients and "Great Coach" Effect Coefficients**

| Coefficient | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|
| Gold | 1.18 | 0.82 | 1.09 | 1.19 | 0.12 | 1.79 |
| Silver | 2.35 | 1.63 | -0.61 | 1.96 | -3.73 | 1.57 |
| Bronze | 1.60 | 1.11 | 0.80 | 1.61 | -0.09 | 0.03 |

We can see that during the coaching years, the "Great Coach" effect has different impacts on gold, silver, and bronze medals. It has a very significant effect on gold medals, a relatively important effect on silver medals, but a very small effect on bronze medals.

### 6.2.2 Quantifying the Contribution of XGBoost Using SHAP

SHAP (SHapley Additive exPlanations) is a model-agnostic method for additive feature attribution. It uses Shapley values from game theory to identify the importance of feature variables. SHAP provides a feasible way to explain machine learning models that are often considered "black boxes," addressing the lack of interpretability in these models while leveraging their predictive accuracy.

The total prediction contribution of a model can be expressed as

$$g(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) I(x_i) \tag{21}$$

$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[ f_{(S \cup i)}(x_{(S \cup i)}) - f_S(x_S) \right] \tag{22}$$

In Equation 21, $x$ represents the $M$-dimensional feature variables, $I(x_i)$ is a binary indicator variable, $g(x)$ is the prediction result, $\phi_0$ is the mean prediction, and $\phi_i(x)$ is the marginal contribution to the prediction.

We use the XGBoost model to calculate the Shapley values to explain the feature importance. The results are shown in Table 11.

**Table 11: Shapley Values Results**

| Sample | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|
| Year 1 | -0.62 | 0.00 | -0.03 | -0.01 | -0.05 | -0.36 |
| Year 2 | 0.25 | 0.00 | 0.04 | -0.01 | -0.09 | 1.74 |
| … | … | … | … | … | … | … |
| Mean (Absolute) | 0.62 | 0.00 | 0.18 | 0.03 | 0.08 | 0.60 |

We can see that the "Great Coach" effect plays a very significant role in predicting medal scores. Traditional advantages also have a noticeable impact, which is consistent with the regression analysis. However, the influence of other features is identified as relatively small.

## 6.3 Investment in the "Great Coach" Effect

We will examine the United States, China, and Romania to determine in which sports each country should invest in the "Great Coach" effect. First, we predict the gold, silver, and bronze medal scores using the regression model described earlier, without considering the "Great Coach" effect. Then, we add the "Great Coach" effect, assuming that we can ignore the differences between different "great coaches" across different sports. We can use the average "Great Coach" effect from Lang Ping and Béla Károlyi as a universal constant.

If the addition of the "Great Coach" effect results in a significant increase in the gold and silver medal scores for a particular sport, we recommend investing in a "great coach" for that sport. Since the "Great Coach" effect has a smaller impact on bronze medals, we will focus only on gold and silver medals.

**Table 12: Medal Score Performance After Investing in the "Great Coach" Effect**

| Country | Sport | Gold Medal Increase Rate | Increase in Gold Medals | Silver Medal Increase Rate | Increase in Silver Medals |
|---|---|---|---|---|---|
| USA | Skateboarding | 21.3% | 1 | 17.2% | 1 |
| | Canoeing | 11.3% | 0 | 14.3% | 1 |
| China | Weightlifting | 27.4% | 2 | 6.9% | 0 |
| | Athletics | 11.4% | 0 | 9.8% | 1 |
| | Judo | 54.1% | 3 | 42.1% | 2 |
| Romania | Athletics | 1.1% | 0 | 0.07% | 0 |
| | Rowing | 10.8% | 0 | 9.4% | 0 |

We can see that for the United States, only Skateboarding and Canoeing show significant advantages after investing in the "Great Coach" effect. Weightlifting also shows a notable increase, but it is not listed as a primary sport in the table. For China, investing in Judo is highly recommended, as it results in a substantial increase in medal counts. In contrast, Romania does not have any sports where the "Great Coach" effect leads to significant changes in medal scores, even after adding the effect.

# 7 Other Original Insights

## 7.1 Do the Most Advantageous Sports for Each Country Share Common Characteristics?

First, we calculate the sport advantage coefficient for each country and identify the most advantageous sport for each country. This most advantageous sport can be considered a characteristic of the country's strong events. We then rank these most advantageous sports in descending order across all countries and plot the top 50 countries on a world map.

From this, we can observe the following characteristics:

- In North America, including Canada, the United States, and Mexico, the most advantageous sports show consistency.
- The top-performing countries in the Olympics, such as the United States, Canada, China, and Australia, primarily excel in Aquatics (indicated by the blue areas on the map).
- Many coastal countries have Aquatics as their most advantageous sport.
- he most advantageous sports for countries show clear regional patterns. For example, Somalia and Kenya, as well as Norway, Belarus, and the Czech Republic, share similar dominant sports.

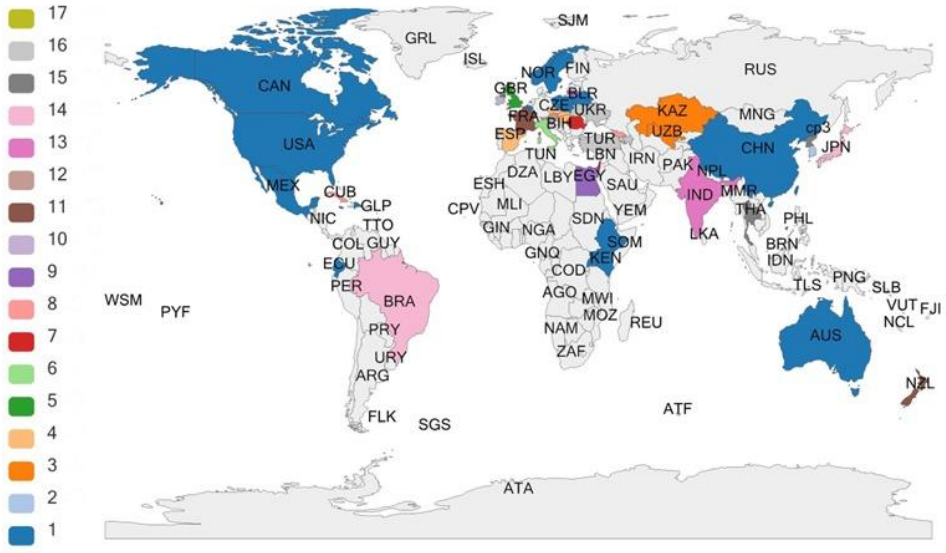


**Figure 13: Advantageous Sports of Various Countries**

## 7.2 Are There Connections Between Different Sports?

Next, we use the sport advantage coefficients of each country to perform sample clustering. First, we remove all countries that do not have a significant advantage in

any sport. These countries, which generally lack a strong sport, are grouped into a single category.

For the remaining countries, we use the K-means clustering algorithm with a cluster count of 2. We visualize the countries on a world map, with countries in the same cluster having the same color.

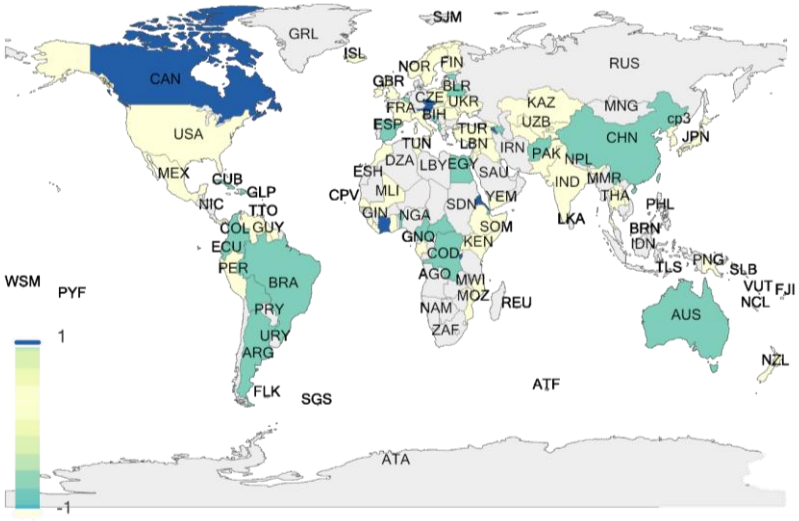


**Figure 14: Distribution of the same cluster on the world map**

We can observe that the distribution of countries' strong sports is fragmented. These sports are clustered in different regions around the world, forming "clusters." Neighboring countries often have similar strong sports.

Next, we perform a correlation analysis for each sport within each cluster.

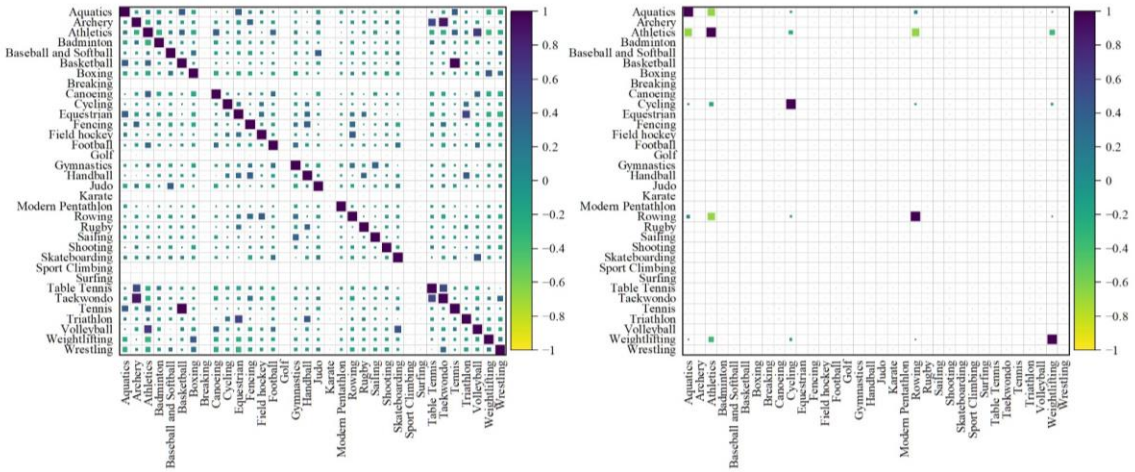


**Figure 15: Correlation Heatmaps (Left: Cluster 1, Right: Cluster 2)**

In Cluster 1, basketball and tennis show a high correlation. Archery also shows a high correlation with taekwondo and table tennis. Strongly correlated sports indicate that when one sport is a country's advantageous sport, the other is likely to be an advantageous sport for that country as well.

In Cluster 2, athletics shows a strong negative correlation with aquatics and archery. This indicates that when athletics is a country's advantageous sport, aquatics and archery are unlikely to be its advantageous sports.

# 8 Sensitivity Analysis

Throughout our work, the most important foundational model is the multi-discipline combination prediction model, where each feature plays a crucial role.

$$y_i(t) = \sum_{i=1}^{6} a_i x_i + \beta_1 y(t-1) + \beta_2 (t-2) + \beta_3 y(t-3) + \beta_0 \tag{23}$$

We will use the Swimming discipline of the United States in the 2024 Olympics as an example.

In the already fitted regression model, we vary the feature coefficient of one discipline at a time by ±10%, and uniformly sample 100 data points.

We can observe that the changes in the feature coefficients result in different levels of medal count fluctuations. The order of impact is as follows: star athlete effect > traditional discipline advantage > number of participants in the discipline > alignment of disciplines > gender ratio in the discipline > potential advantage of the discipline. Star athletes can significantly drive large fluctuations in medal counts, while the potential advantage of the discipline has a relatively weaker impact on medal count fluctuations.

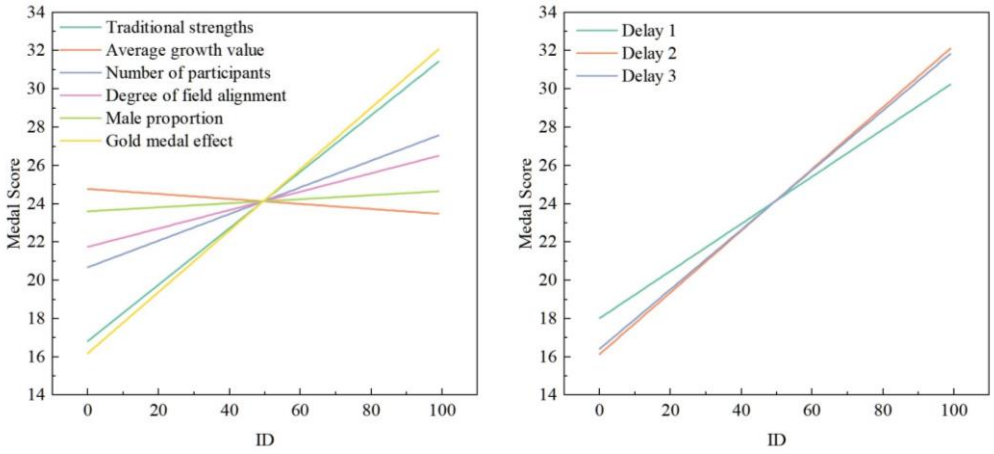


**Figure 16: Sensitivity Analysis of Feature Values (Left: Six Feature Coefficients, Right: Three Lag Term Coefficients)**

For the three lag term coefficients, we find that their fluctuations have very similar impacts. Particularly, the second and third lag terms are almost identical. This also reflects the difference between the first lag term and the other two. The recent historical data (first lag term) has a more stable influence, while the influence of more distant historical data (second and third lag terms) is more volatile.

# 9 Model Evaluation and Further Discussion

## 9.1 Strengths

- The multi-discipline combination prediction model we developed considers both the feature information for multiple years to be predicted and the impact of historical data on those years. The model can accurately predict the number of gold, silver, and bronze medals for each discipline and even make reasonable predictions for individual events within those disciplines.

- While achieving its predictive goals, the model is highly interpretable, allowing us to explain the importance of each feature.
- The model is highly flexible and can be easily adjusted to include different features based on specific circumstances. This makes it easy to study additional factors, such as the "Great Coach" effect.

### 9.2 Weaknesses and Further discussion

- The main limitation of the model is that combining regression fitting with coefficient prediction often requires a significant amount of computational resources.
- During coefficient prediction, the relatively small number of sample points can lead to the model easily deviating from the controllable error range.

Therefore, further improvements to the model should focus on enhancing the accuracy of coefficient prediction. This can be achieved by incorporating data from more Olympic Games or by using records from other sporting events to increase the number of training samples.

Additionally, methods such as machine learning can be employed to improve prediction accuracy. For example, as the number of sample points increases, ARIMA can be used to fit the linear components, while neural networks like LSTM can be used to fit the nonlinear components, thereby enhancing the overall prediction accuracy.

## 10 Conclusion

First, we build a multi-Discipline combination prediction model based on feature construction. This model accurately predicts the gold, silver, and bronze medals for each country. We use this model to make reasonable predictions for the medal tally of the 2028 Los Angeles Summer Olympics. The top seven countries expected to perform well in the 2028 Olympics are the United States, China, Japan, the United Kingdom, Australia, France, and South Korea. The three countries with the most significant improvement are Slovakia, Fiji, and Turkmenistan. The three countries with the worst performance are Hong Kong, the Philippines, and Portugal.

Next, we predict that four countries will win their first medals in 2028: Papua New Guinea, Haiti, Seychelles, and Nepal. We also analyze the characteristics of countries that won their first medals in past Olympics. These countries often gain experience through multiple participations. We verify the characteristics of the four predicted first-time medal-winning countries and find they fit this pattern, validating the model's predictions.

Then, we explore the impact of the number and type of events on medal counts. We construct a multiple linear regression model to explain this impact. Additionally, we build a simpler linear regression model to find the most direct linear relationship. We also calculate the advantage degree of each sport for a country to measure its importance. Furthermore, we analyze the effect of host country event selection on Olympic medal outcomes and discover a surprising host country effect.

Moreover, we visualize the medal score changes before and after the coaching of "great coaches." This reveals the potential existence of a "great coach" effect. We attribute this effect using both a modified regression model and an XGBoost model with SHAP interpretability. We quantify its impact on medal scores. The effect is significant for gold and silver medals but weaker for bronze medals.

We then use the quantified "great coach" effect to determine the suitable sports for investing in "great coaches" for the United States, China, and Romania. We also provide the increase rates of gold and silver medal scores after adding the "great coach" effect, to estimate its impact on medal counts.

Additionally, we visualize the advantage sports of various countries on a world map based on their sport advantage coefficients. We clearly observe that North America shows consistency in its advantage sports. Countries like the United States and China, which rank high in the Olympics, have Aquatics as their main advantage sport, and they are all coastal countries. The advantage sports of different countries show clear regional patterns.

Furthermore, we use the K-means clustering algorithm to classify countries into three categories based on their sport advantage coefficients. We visualize the clustered countries on a world map and find that the advantage sports of these countries show significant "clustered" patterns. We then conduct a correlation analysis of the sports within each cluster and find, for example, that basketball and tennis show a very significant correlation.

Finally, we perform a sensitivity analysis on the multi-Discipline combination prediction model. We evaluate the impact of each feature on the model's medal count predictions and test the model's stability.

# References

[1] Tian, H., He, Y., Wang, M., Li, J., Yu, P., Qi, S., & Tian, Y. (2021). Medal prediction and participation strategy for Chinese athletes in the 2022 Beijing Winter Olympics: Analysis based on the home advantage effect. *Journal of Sports Science*, 3-13 + 22. https://doi.org/10.16469/j.css.202102001

[2] Shi, H., Zhang, D., & Zhang, Y. (2024). Can Olympic medals be predicted? — A perspective from explainable machine learning. *Journal of Shanghai University of Sport*, 26-36. https://doi.org/10.16099/j.sus.2023.10.27.0002

[3] Wang, G., Xue, E., & Tang, X. (2010). Research on predicting the number of medals in large international comprehensive sports events: A case study of the Beijing Olympics. *Journal of Tianjin University of Sport*, 86-90. https://doi.org/10.13297/j.cnki.issn1005-0000.2010.01.007

[4] Forrest, D., Sanz, I., & Tena, J. D. (2010). Forecasting national team medal totals at the Summer Olympic Games. *International Journal of Forecasting*, 26(3), 576-588.

[5] Schlembach, C., Schmidt, S. L., Schreyer, D., & Wunderlich, L. (2020). Forecasting the Olympic medal distribution during a pandemic: A socio-economic machine learning model. [Unpublished manuscript].

[6] Fulton, T. J., Baranauskas, M. N., & Chapman, R. F. (2022). World championship and Olympic Games experience influences future medal performance in track-and-field athletes. *International Journal of Sports Physiology and Performance*, 17(1), 111-114.